基于节点地位和相似性的社交网络边符号预测*

卢志刚，叶美丽

(上海海事大学 经济管理学院，上海 201306)

摘　要：边符号预测即根据网络拓扑结构挖掘符号相关的隐含信息，旨在揭示用户之间的潜在关系。节点地位和相似性能够较好地体现边符号属性，为改善预测效果提供了理论基础。通过探究二者与边符号属性之间的强相关性，建立符号预测模型。首先，利用排序算法 Prestige 评估用户节点的社会地位，同时使用余弦相似度表示用户的社交偏好；然后，在逻辑回归学习模型的基础上融合二者建立边符号预测模型 LR-SN；最后，在模型的训练过程中采用随机梯度上升算法优化求解。三个真实网络数据集的实验结果表明，相比于现有基准方法，LR-SN 模型的符号预测准确率显著提高且具有一定的推广性，说明通过融合局部信息与全局信息能够进一步改善预测效果。

关键词：边符号预测；节点地位；节点相似性；逻辑回归；随机梯度上升算法

中图分类号：TP391　　doi: 10.19734/j.issn.1001-3695.2018.07.0516

Social network edge sign prediction based on node status and similarity

Lu Zhigang, Ye Meili

(*College of Economics & Management, Shanghai Maritime University, Shanghai* 201306，China)

Abstract: The edge sign prediction is to mine the sign-related implicit information according to the network topology, aiming to reveal the potential relationship between users. Node status and similarity can better represent sign attributes of edges, providing a theoretical basis for improving the prediction effect. By investigating the strong correlation between the two theories and the sign attributes of the edges, a sign prediction model is established. Firstly, use prestige evaluate the social status of user nodes. At the same time, cosine similarity can represent the user's social preferences. Then, both sides are combined based on the logistic regression learning model to establish the edge sign prediction model LR-SN. Finally, a random gradient ascent algorithm will optimize the model during training. The experimental results of three real network datasets show that compared with the existing baseline methods, the accuracy of sign prediction of LR-SN model is significantly improved and has certain generalization, indicating that the fusion of local information and global information can further improve the prediction effect.

Key words: edge sign prediction; node status; node similarity; logistic regression; random gradient ascent algorithm

0　引言

社交网络是人们进行意见交流与信息共享的虚拟空间，其允许用户将与之有关联的个体标记为朋友或者敌人关系，对他人的言论及观点提出赞同或者反对意见。因此可以将社交网络描述为边具有正或者负符号属性的有向网络，其中正向边表示两个用户之间具有朋友、信任、喜欢等积极关系；而负向边则表示两个用户之间具有敌对、怀疑、厌恶等消极关系。社交网络中的边符号预测即通过提取网络结构信息和用户关系数据预测未知的边符号，它揭示了用户之间的潜在关系如朋友，陌生人，敌人等。

边符号预测在机器学习、大数据分析与决策等领域具有重要的研究意义。探究边的符号属性有助于理解网络基本结构特征[1]，解决个性化推荐[2]、舆情分析[3]、异常用户检测[4]等问题。本文深入研究边的符号属性，提出一种高效的边符号预测模型，并在 Epinion、Slashdot、Wikipedia 数据集上建立多组实验，结果证明了该模型在符号预测方面的有效性。主要贡献如下：

a）提出两个有关符号属性的量化策略，分别量化节点地位以及相似性。

b）在逻辑回归学习模型的基础上，融合节点地位和相似性建立边符号预测模型 LR-SN，其中节点地位从全局角度量化符号属性相关特征，节点相似性从局部角度体现符号属性。

c）为证明 LR-SN 模型的有效性，在 Epinion、Slashdot、Wikipedia 数据集上建立多组实验，并详细阐述不同量化策略对符号预测准确率的影响。

1　相关性研究

社交网络边符号的研究起源于社会心理学，起初由 Heider 等人[5]从心理学角度出发，探讨了人际交往中正关系与负关系的相互作用模式。随后 Cartwright 和 Harary 等人[6]以图论的语言将社交网络描述为边具有正负符号属性的有向网络。随着复杂网络的兴起，社交网络中的边符号预测问题逐渐成为研究的热点。

目前，有关边符号预测的方法大致分为两类：考虑局部特征的方法和考虑全局特征的方法。考虑局部特征的方法仅仅利用节点的领域特征如节点出入度[7]，共同邻居数量[7]，节点相似性[8-9]等进行边符号预测。而考虑全局特征的方法扩大了特征提取的范围，从全局角度量化网络的不平衡程度，

收稿日期：2018-07-07；**修回日期：**2018-08-20　**基金项目：**上海市自然科学基金资助项目（18ZR1416900）

作者简介：卢志刚（1973-），男，教授，博士，主要研究方向为大数据分析与决策、商务智能、供应链管理（963620627@qq.com）；叶美丽（1994-），女，硕士研究生，主要研究方向为大数据分析与决策、数据挖掘、商务智能.

一般采取扩展的结构平衡理论[10]、上下文信息[11, 12]、节点排序[13]等措施对边符号进行预测。Leskovec 等人[7]对符号预测问题进行了形式化定义，其通过提取两类网络结构信息：节点邻域特征以及基于社会学理论的 16 种三元组关系模式，然后利用逻辑回归训练特征实现了边符号预测。Chiang 等人[10]提出利用扩展的结构平衡有序长环对边符号进行预测，实验表明当环的长度由 3 递增到 5 时，预测准确率有效提高。该方法实现了对 Leskovec 等人局部度量方法的扩展。Symeonidis 等人[9]通过定义同一簇之间的相似性与不同簇之间的相似性，然后利用推荐算法实现了边符号预测。Shahriari 和 Jalili 等人[13]提出将排序算法引入到特征值的计算当中，其首先利用各类排序算法对网络中的节点进行排序，然后基于该节点排序值计算特征，该方法实现了从全局角度体现边符号属性。

网络中的局部信息与全局信息联系密切，但边符号预测仅使用二者之一是不够全面的。针对以上问题，本文在逻辑回归学习模型的基础上，通过引入节点地位与相似性两种量化策略，实现局部信息与全局信息的融合，从而解决由于网络稀疏，局部特征利用不足导致的预测准确率较低等问题。

2　问题形式化描述

将社交网络用一个有向网络图来表示，记为 $G(V,E,S)$，其中 $V=\{1,2,\ldots,n\}$ 代表社交网络中节点用户的集合，$E=\{1,2,\ldots,m\}$ 代表网络中节点用户之间关系的边集合，$S=\{1,-1,0\}$ 代表边符号。$i,j\in V$，$e(i,j)\in E$，$s(i,j)\in S$，其中 $s(i,j)=1$ 代表“+”，表示节点 i 和节点 j 之间具有信任，合作，友好，支持等积极关系；$s(i,j)=-1$ 代表“-”，表示两个节点之间具有不信任、敌对、讨厌、否决等消极关系；$s(i,j)=0$ 表示节点 i 与节点 j 之间的互动关系未知。

利用上述符号与定义，对社交网络下的边符号预测问题作如下定义：设计一个符号预测框架 f，通过提取网络结构信息结合已知的用户之间的关系数据来预测网络中的边符号，即给定一个社交网络 $G(V,E,S)$，通过学习符号预测框架 f，预测未知的正负关系 $s(i,j)$，即

$$f=G(V,E,S)\rightarrow s(i,j) \tag{1}$$

3　节点地位和相似性量化

实证发现，边符号属性与节点的地位和相似性息息相关[7~9]，二者在一定程度上能够体现出用户在网络中的受欢迎程度以及社交偏好。因此，利用节点地位和相似性来设计符号预测框架 f，并在此基础上提出两个有关符号属性的量化策略。

3.1　节点地位

利用网络中节点的地位差异能够标记边的符号[14]。具体而言，由 i 指向 j 的正边代表 j 的地位比 i 高，由 i 指向 j 的负边代表 j 的地位低于 i，这种地位高低关系具有传递性[15]。网络拓扑结构可以用来评估节点用户的社会地位。在社交网络中，正向边入度有助于提高节点的社会地位，负向边入度相反会降低节点的社会地位，由此提出引入考虑边符号属性的节点地位量化策略-Prestige 来评估用户的社会地位。

Prestige [12]仅仅考虑网络结构中节点的正负向边入度。如果一个节点收到很多来自其他节点的正向边，说明该节点在网络中具有较高的地位和威望。相反的，如果该节点收到很多来自其他节点的负向边，说明它在网络中的地位和信誉较低。节点 i 的 Prestige 值（Pr_i）计算如下：

$$\mathrm{Pr}_i=\frac{|IN_i^+|-|IN_i^-|}{|IN_i^+|+|IN_i^-|} \tag{2}$$

其中：$|IN_i^+|$，$|IN_i^-|$ 分别表示节点 i 的正向边入度以及负向边入度。Prestige 值越高，该节点具有越高的威望和地位，其在网络中越容易被信任。相反，若 Prestige 值较低，则该节点较难被其他节点信任。

3.2　节点相似性

具有类似偏好的用户倾向给予社交网络中的边类似的符号[8]。节点相似性能够大致反映网络中用户的社交偏好。对于待预测边符号的相关节点，通过计算源节点与目的节点的邻居节点之间的平均相似度，能够推断出源节点给予目的节点正向边或者负向边的可能性。直观理解，如果用户 i 与给予用户 j 正向边的用户均具有较高的相似度，那么用户 i 则有很大的可能性给予用户 j 正向边。相反，如果用户 i 与给予用户 j 负向边的用户具有较高的相似度，则用户 i 有很大的可能性给予用户 j 负向边。用 $r_{i,p}$ 表示节点 i 指向节点 p 的边符号，I 表示节点 i 和节点 k 的邻居节点，使用余弦相似度来计算用户节点之间的相似性，$Sim(i,k)$ 的定义如下：

$$Sim(i,k)=\frac{\sum_{p\in I}r_{i,p}r_{k,p}}{\sqrt{\sum_{p\in I}r_{i,p}^2\sum_{p\in I}r_{k,p}^2}} \tag{3}$$

基于已知的社交网络拓扑信息，通过节点相似性量化用户 i 对用户 j 给予正向边的可能性 $S^+(i,j)$：

$$S^+(i,j)=\frac{\sum_{k\in W^+}Sim(i,k)}{|W^+|} \tag{4}$$

相反，用户 i 对用户 j 给予负向边的可能性 $S^-(i,j)$：

$$S^-(i,j)=\frac{\sum_{k\in W^-}Sim(i,k)}{|W^-|} \tag{5}$$

其中：W^+，W^- 分别表示与节点 j 产生正向边以及负向边的节点集合，$|\bullet|$ 表示集合的数量。

4　边符号预测模型及优化算法

以节点地位和相似性作为模型建立的基准，在基于逻辑回归的边符号预测方法基础上，分别建立相应量化策略的边符号预测模型。

4.1　基于逻辑回归的边符号预测 LR

逻辑回归是一种有监督的统计学习方法，其将社交网络中的正负关系预测视为二元分类问题。运用该算法进行符号预测首先要构建边符号属性相关的特征集，然后将特征集作为输入，训练分类器实现正负关系预测，其具体形式如下所示：

$$h_\theta(x)=g(\theta^T x)=\frac{1}{1+e^{-\theta^T x}} \tag{6}$$

$$P(y=1\,|\,x;\theta)=h_\theta(x) \tag{7}$$

$$P(y=0\,|\,x;\theta)=1-h_\theta(x) \tag{8}$$

其中 x=（$x_0,x_1,x_2\ldots$）表示从社交网络中提取的特征向量，一般要求该向量在一定程度上体现边的符号属性。θ=（$\theta_0,\theta_1,\theta_2\ldots$）表示赋予每个特征的权重向量，其通过最大似然法估计。

y 表示要预测的边，有 0,1 两种取值。当 y=1 时预测边为正号，y=0 时预测边为负号。$P(y=1\,|\,x;\theta)$ 表示当预测边 y=1 时的概率。一般情况下，当 $P(y=1\,|\,x;\theta)>0.5$ 时，预测边 y 的数值为 1 即正号，当 $P(y=1\,|\,x;\theta)<0.5$ 时，预测边 y 的数值为 0 即负

号。

4.2 基于节点地位的边符号预测模型 LR-S

为了模拟节点地位用于边符号预测，尝试利用节点地位与边符号属性之间的强关联性建立基于节点地位的边符号预测模型 LR-S。同时，为了强调节点地位在边符号预测过程中的重要性，以其为基准提出了四种节点地位量化特征。这些特征实现了从全局角度量化每个用户节点的社会地位以及乐观程度，较大程度反映了边的符号属性。

节点地位量化特征包括 Rep_i，Opt_i，Rep_j，Opt_j 等四个特征值，分别表示节点 i 的信誉值，乐观值以及节点 j 的信誉值，乐观值。在特征值的计算过程中，首先使用节点地位量化策略-Prestige 评估每个用户节点的社会地位，然后基于该算法得到的 Prestige 值来计算边符号预测相关节点的信誉值与乐观值。

节点 i 的信誉值，乐观值被定义为如下：

$$Rep_i = \frac{\sum_{k\in IN_i^+} \Pr_k - \sum_{k\in IN_i^-} \Pr_k}{\sum_{k\in IN_i^+} \Pr_k + \sum_{k\in IN_i^-} \Pr_k} \tag{9}$$

$$Opt_i = \frac{\sum_{k\in OUT_i^+} \Pr_k - \sum_{k\in OUT_i^-} \Pr_k}{\sum_{k\in OUT_i^+} \Pr_k + \sum_{k\in OUT_i^-} \Pr_k} \tag{10}$$

节点 j 的信誉值，乐观值定义如下：

$$Rep_j = \frac{\sum_{k\in IN_j^+} \Pr_k - \sum_{k\in IN_j^-} \Pr_k}{\sum_{k\in IN_j^+} \Pr_k + \sum_{k\in IN_j^-} \Pr_k} \tag{11}$$

$$Opt_j = \frac{\sum_{k\in OUT_j^+} \Pr_k - \sum_{k\in OUT_j^-} \Pr_k}{\sum_{k\in OUT_j^+} \Pr_k + \sum_{k\in OUT_j^-} \Pr_k} \tag{12}$$

其中：$\Pr_k$ 是节点 k 的排序分数，该值通过式（2）计算得出。IN_i^+, IN_i^- 分别表示对节点 i 给出正向边以及负向边的节点集合；OUT_i^+, OUT_i^- 分别表示收到节点 i 给出的正向边以及负向边的节点集合。类似的，$IN_j^+, IN_j^-, OUT_j^+, OUT_j^-$ 表述同上。

节点的信誉值体现了该节点在网络中的受欢迎程度，它能够衡量其在朋友圈的接受度以及在社会中的号召力。信誉值越高的节点，其地位及影响力越高，网络中其他节点给予该节点正向边的概率也就越大。相反，乐观值体现了该节点对网络中其他节点给予积极友好互动关系的倾向，这在一定程度上说明了该节点的自身性格—“乐观友好型”。乐观值越高，该节点给予其他节点正号边的可能性越大。

信誉值与乐观值之间的逻辑关系如图 1 所示，信誉值仅考虑与节点 i 的入边相关的节点集合，并计算该集合中所有节点的 Prestige 值。相反，乐观值仅考虑与节点 i 的出边相关的节点集合以及该集合中所有节点的 Prestige 值。

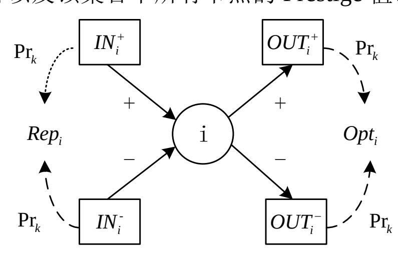


图 1 节点地位量化特征示意图

Fig. 1 Status theory quantitative feature diagram

在基于逻辑回归的符号预测方法基础上，通过引入四种节点地位量化特征作为特征集输入，进而实现模型扩展。由此，LR-S 模型可以定义为如下：

$$P(y=1\,|\,x;\theta) = h_\theta(x) = \frac{1}{1+e^{-(\theta_0+\theta_1 Rep_i+\theta_2 Rep_j+\theta_3 Opt_i+\theta_4 Opt_j)}} \tag{13}$$

4.3 基于节点地位和相似性的边符号预测模型 LR-SN

节点地位虽然能够从全局角度量化符号属性相关特征，但仍旧存在局限。其一，当两个用户的社会地位差距不大时，其预测准确率将受到影响；其二，节点地位重点关注网络中的全局信息，而忽视了局部信息。实证发现，当网络聚类系数较高时，局部信息反而能够获得较全局信息更高的预测准确率。因此，当建立基于节点地位的边符号预测模型时，不仅要将两个用户之间的社会地位差异考虑进来，还需结合体现局部信息的节点相似性来克服其在符号预测方面的局限性。由此提出建立基于节点地位和相似性的边符号预测模型 LR-SN。相比于 LR-S，该模型在其基础上又创新的提出了四种节点相似性量化特征。这些特征实现了从局部角度量化每个用户的属性及偏好，并进一步体现了边符号属性。

节点相似性量化特征，包括源节点正相似度，源节点负相似度，目的节点正相似度以及目的节点负相似度，具体定义如下：

a）源节点正相似度 $S_{out}^+(i,j)$.节点 i 和给 j 正向边的节点之间的平均相似度。$S_{out}^+(i,j)$ 的值越高，意味着节点 i 指向节点 j 的边是正号的可能性越大。$S_{out}^+(i,j)$ 的计算如下：

$$S_{out}^+(i,j) = \frac{\sum_{k\in W_{out}^+} Sim_{out}(i,k)}{|W_{out}^+|} \tag{14}$$

其中：W_{out}^+ 是对节点 j 给出正向边的节点集合，$Sim_{out}(i,k)$ 是节点 i 与给 j 正向边的节点 k 之间的相似度，具体公式如下：

$$Sim_{out}(i,k) = \frac{\sum_{p\in I_{out}} r_{i,p} r_{k,p}}{\sqrt{\sum_{p\in I_{out}} r_{i,p}^2 \sum_{p\in I_{out}} r_{k,p}^2}} \tag{15}$$

其中：$r_{i,p}$ 和 $r_{k,p}$ 分别是节点 i 与 k 分别指向节点 p 的边符号，I_{out} 是节点 i 与 k 均给出边的节点集合。

b）源节点负相似度 $S_{out}^-(i,j)$.节点 i 和给 j 负向边的节点之间的平均相似度。$S_{out}^-(i,j)$ 的值越高，节点 i 指向节点 j 的边是负号的概率越大。$S_{out}^-(i,j)$ 的公式定义为如下：

$$S_{out}^-(i,j) = \frac{\sum_{k\in W_{out}^-} Sim_{out}(i,k)}{|W_{out}^-|} \tag{16}$$

其中：W_{out}^- 是给节点 j 负向边的节点集合，$Sim_{out}(i,k)$ 是节点 i 和 k 之间的相似度，由式（15）计算。

c)目的节点正相似度 $S_{in}^+(j,i)$.节点 j 和从 i 接收正向边的节点之间的平均相似度。$S_{in}^+(j,i)$ 的值越高，节点 j 收到来自节点 i 的正向边的概率越大。$S_{in}^+(j,i)$ 的公式定义为如下：

$$S_{in}^+(j,i) = \frac{\sum_{k\in W_{in}^+} Sim_{in}(j,k)}{|W_{in}^+|} \tag{17}$$

其中：W_{in}^+ 是节点 i 给出正向边的节点集合，$Sim_{in}(j,k)$ 是节点 j 与 k 之间的相似度。$Sim_{in}(j,k)$ 的计算公式如下：

$$Sim_{in}(j,k) = \frac{\sum_{p\in I_{in}} r_{p,j} r_{p,k}}{\sqrt{\sum_{p\in I_{in}} r_{p,j}^2 \sum_{p\in I_{in}} r_{p,k}^2}} \tag{18}$$

其中：$r_{p,j}$ 和 $r_{p,k}$ 分别代表从节点 p 指向节点 j 与 k 的边符号，I_{in} 是节点 j 与 k 均从中接收边的节点集合。

d)目的节点负相似度 $S_{in}^-(j,i)$.节点 j 和从 i 接收负向边的节点之间的平均相似度。$S_{in}^-(j,i)$ 的值越高，节点 j 将有更大概率收到来自节点 i 的负向边。$S_{in}^-(j,i)$ 的公式定义为如下：

$$S_{in}^{-}(j,i)=\frac{\sum_{k\in W_{in}^{-}} Sim_{in}(j,k)}{|W_{in}^{-}|} \tag{19}$$

其中：W_{in}^{-} 是节点 i 给出负向边的节点集合，$Sim_{in}(j,k)$ 是节点 j 与 k 之间的相似度，由式（18）计算。

四种节点相似性量化特征之间的逻辑关系如图 2 所示。在基于逻辑回归的边符号预测基础上，融合节点地位和相似性建立边符号预测模型。节点地位量化特征从全局角度反映节点的社会地位以及乐观程度，节点相似性量化特征从局部角度反映节点属性及偏好。通过合并二者作为特征集输入，实现模型的进一步扩展。LR-SN 模型定义如下：

$$P(y=1|x;\theta)=h_\theta(x)=\frac{1}{1+e^{-(\theta_0+\theta_1 Rep_i+\theta_2 Rep_j+\theta_3 Opt_i+\theta_4 Opt_j+\theta_5 S_{out}^{+}(i,j)+\theta_6 S_{out}^{-}(i,j)+\theta_7 S_{in}^{+}(j,i)+\theta_8 S_{in}^{-}(j,i))}} \tag{20}$$

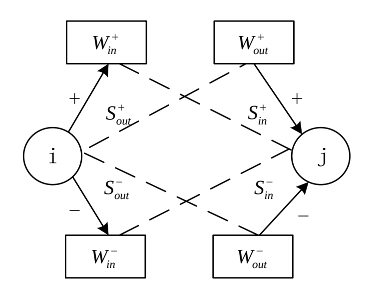


图 2　节点相似性量化特征示意图

Fig. 2　Node similarity quantization feature diagram

4.4　优化算法

针对节点地位与相似性分别建立基于节点地位的边符号预测模型 LR-S 以及结合节点地位和相似性的边符号预测模型 LR-SN，其中节点地位从全局角度量化符号属性相关特征，节点相似性从局部角度体现符号属性。通过合并二者作为特征集输入，采用式（6）的逻辑回归方法建模边符号预测问题。在模型的训练过程中，使用随机梯度上升算法优化求解。该算法在每次迭代过程中仅根据随机选择的一个样本来更新权重向量，相比于传统的梯度上升算法，计算量大大降低，训练速度更快。

将式（7）（8）合并得到一个样本的代价函数，形式如下：

$$Cost(h_\theta(x),y)=h_\theta(x)^{(y)}(1-h_\theta(x))^{(1-y)} \tag{21}$$

代价函数被用来估计预测值与实际值之间的误差。给定一个样本，即可通过代价函数求出该样本所属类别的概率。假定每个样本之间彼此独立，那么整个样本集发生的概率即为所有样本发生概率的乘积。为了方便求解,对该式对数化，则得到整个样本集的代价函数，形式如下：

$$J(\theta)=\sum_{i=1}^{m}[y^{(i)}\log(h_\theta(x^{(i)}))+(1-y^{(i)})\log(1-h_\theta(x^{(i)}))] \tag{22}$$

其中：m 为样本总数，$y^{(i)}$ 表示第 i 个样本的符号，$x^{(i)}$ 表示第 i 个样本。

为了求出使得代价函数最大时的 θ 值，使用随机梯度上升算法优化求解，其迭代公式如下：

$$\alpha=\frac{10}{1+j+k}+0.01 \tag{23}$$

$$\frac{\partial J(\theta)}{\partial\theta_j}=(y^{(i)}-h_\theta(x^{(i)}))x_j^{(i)}\quad j\in\{0,1,2,\ldots,8\} \tag{24}$$

$$\theta_j:=\theta_j+\alpha(y^{(i)}-h_\theta(x^{(i)}))x_j^{(i)} \tag{25}$$

其中：k 为迭代次数，α 为步长，也就是学习速率，用来控制更新的幅度。在训练过程中步长 α 的选取至关重要，若 α 取值过大将导致结果不收敛甚至出现发散等现象；若 α 取值过小则会使得迭代次数增多，收敛速度缓慢。为了使得算法稳步进行，令 α 在每次迭代过程中减少 $1/j+k$，从而有效解决固定步长引起的在最优值附近波动等问题。具体步骤如算法 1 所示。

算法 **1**　LR-SN.

输入：社交网络 $G(V,E,S)$

输出：$s(i,j)$

1. 依据式（2）计算每个节点的 Prestige 值
2. 依据式（9）-式（12）计算节点地位量化特征
3. 依据式（14）-式（19）计算节点相似性量化特征
4. 输入训练样本集
5. while 不收敛 do

计算 $\alpha=\frac{10}{1+j+k}+0.01$

计算 $\frac{\partial J(\theta)}{\partial\theta_j}$；

更新 $\theta_j\leftarrow\theta_j+\alpha\frac{\partial J(\theta)}{\partial\theta_j}$；

6. end while

5　实验分析

5.1　数据集描述

通过在三个真实世界数据集（Epinion，Slashdot，Wikipedia，数据集可以从 Snap 网站上下载）上进行实验来验证所提出模型的有效性。其中 Epinion 是一个在线评论网站，在网络中，人们用 1 和-1 等符号表示他们对彼此的看法。Slashdot 是一个技术新闻网站，用户在网站上可以将对方标记为敌人或者朋友。维基百科是由世界各地的志愿者创建的著名百科全书，其管理人员通过投票选举产生，用户可以通过投票表示赞成或者反对该候选人。表 1 给出了三个真实数据集的统计特征。

表 1　数据集统计信息

Table 1　Data set statistics

属性	Epinion	Slashdot	Wikipedia
节点	131828	77350	7065
边	841372	516575	103561
正边(%)	85%	77%	78.7%
负边(%)	15%	23%	21.3%
平均聚类系数	0.1279	0.0549	0.0691

表 1 的统计结果显示，三个网络中负号边的占比均在 25%以下，网络中负号边的数量远远小于正号边的数量。由心理学和社会学可知，出于礼貌举止或者害怕被报复等心理，人们在社交网络中很少对其他用户表现出反感、讨厌等情绪。此外，三个网络中 Epinion 的平均聚类系数最高，表明其节点分布最密集，其次为 Wikipedia，Slashdot 的网络聚类系数最低。

5.2　实验设置与评价指标

由数据集统计信息可知，三个网络数据集中正号边与负号边分布极不均匀，这将导致符号预测的准确率具有较低的可信度。为此，实验过程中采取随机抽样的方法将数据集分成训练集与测试集两部分，依次随机选择 10%、30%、50%、…、90%的数据集用于训练，剩余的 90%、70%、50%、…、10% 的数据集用于测试。另外，为了保证预测结果的可靠性，将上述实验重复 5 次取平均值。

使用精确度（accuracy）来评价预测算法对边符号的预

测准确率，如式（26）所示。同时，利用混淆矩阵来表示预测结果，如表 2 所示。

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{26}$$

表 2　混淆矩阵

Table 2　Confusion matrix

real vale	prediction	
	positive	negative
Positive	TP(true positive)	FN(flase positive)
Negative	FN(flase positive)	TN(true negative)

5.3　边符号预测模型对比效果

为了证明所提出边符号预测模型的有效性，将 LR-S 模型以及 LR-SN 模型与以下三种基准方法进行比较。

a）Status 方法。该方法是基于节点地位与边符号属性之间的强关联性而定义的，其依照待预测两个用户之间的社会地位差异预测符号。正向地位差距越大，其边符号越有可能是正号；负向地位差距越大，其边符号越有可能是负号。

b）Balance 方法。该方法是基于结构平衡理论[5]与边符号属性之间的强关联性定义的。在进行符号预测时，根据待测边所在三元组的结构平衡性即可推断出边符号。该理论依据朋友的朋友是我的朋友，敌人的朋友是我的敌人[17]等直观认识，判定当三角形拥有奇数个正号边时结构平衡。

c）LR 方法[7]。该方法是基于节点地位与结构平衡理论提出的，主要从网络中提取两类特征。第一类特征为度特征，包括节点 i 的总出度 $d_{out}(i)$ 、正出度 $d^{+}_{out}(i)$ 、负出度 $d^{-}_{out}(i)$ ，节点 j 的总入度 $d_{in}(j)$ 、正入度 $d^{+}_{in}(j)$ 、负入度 $d^{-}_{in}(j)$ ，以及节点 i 和 j 的共同邻居数量 $C(i,j)$ 。第二类特征为三元组特征，包括了待测边所处的十六种不同形式的三元组。将以上两类特征基于逻辑回归方法进行训练，从而实现符号预测。

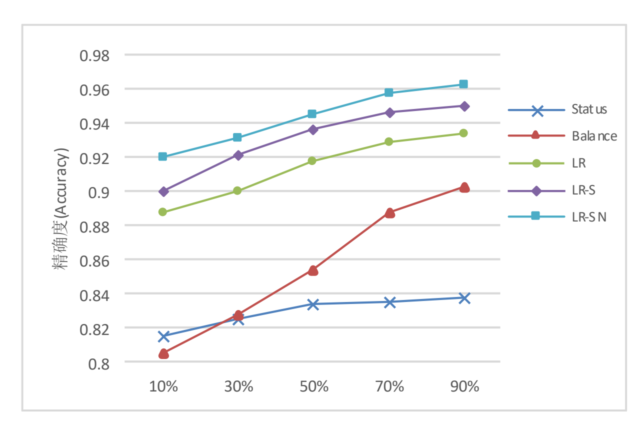
在 Epinion，Slashdot 和 Wikipedia 三大数据集上分别使用上述方法进行符号预测，实验对比结果如图 3 所示。

通过分析图 3（a）（b）（c）可以看出：

a）LR-S 模型和 LR-SN 模型的符号预测精确度均高于其他基本方法，尤其是明显高于 Status 方法和 Balance 方法，说明节点地位和相似性量化策略的有效性。此外，与 LR-S 模型以及 LR 方法相比，LR-SN 模型在实验中表现最好，预测精确度平均高于 LR-S 模型 0.73%，高于 LR 方法 3.32%。说明通过结合节点地位与相似性实现局部信息与全局信息的融合，能够有效提高符号预测准确率。

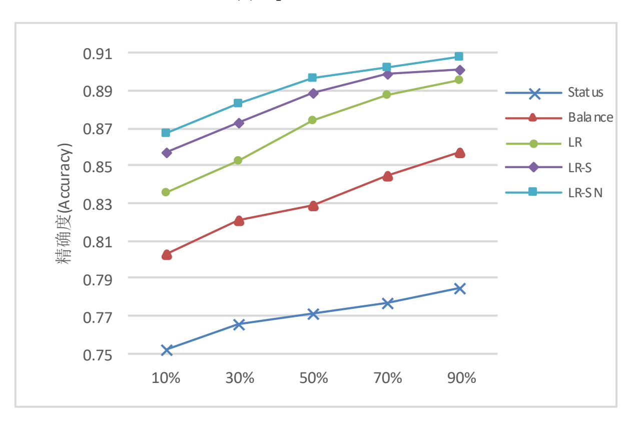
b）从 LR-S 模型和 LR-SN 模型的对比结果可以发现，将考虑局部信息的节点相似性量化特征加入特征集后，预测准确率得到了改善。例如在 Epinion 数据集中，LR-SN 模型在训练集占比为90%时的预测精确度为96.31%，相较于 LR-S 模型提高了 1.29%。在其他两个数据集上，其改善效果没有 Epinion 数据集明显，分析其原因，可能是因为 Slashdot 和 Wikipedia 数据集较为稀疏，网络中能获取的局部信息较少。

c）通过比较 Status 方法和 Balance 方法发现，Balance 方法在三个数据集中的表现均优于 Status 方法，其中 Epinion 数据集的对比效果尤为明显。这在某方面说明了考虑局部信息的方法在符号属性方面的预测效果要优于考虑全局信息的方法，尤其是在网络较为密集的情况下。另外， LR 方法也获得了较好的预测效果，该方法综合考虑了地位理论与结构平衡理论，相较于 Status 方法和 Balance 方法，其预测精确度显著提高，进一步验证了通过融合全局信息和局部信息能够有效提高预测精确度。



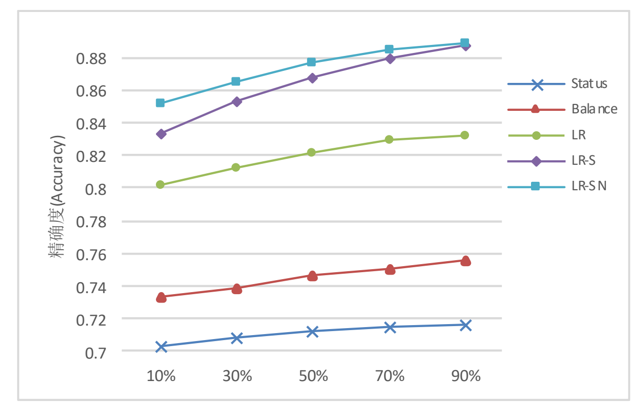
(a) Epinion 数据集

(a) Epinion dataset



(b)Slashdot 数据集

(b) Slashdot dataset



(c)Wikipedia 数据集

(c) Wikipedia dataset

图 3　不同符号预测方法对比结果

Fig. 3　Comparison of different sign prediction methods

此外，实验中还对 LR-S 模型以及 LR-SN 模型的泛化能力进行了测试。运用 3 组数据集训练的两种模型均体现出较好的泛化能力，如表 3、4 所示。

表 3 符号预测模型 lr-s 的泛化能力（训练集占比 90%）

Table 3　Generalization ability of the sign prediction model LR-S (training ratio 90%)

训练集	测试集		
	Epinion	Slashdot	Wikipedia
Epinion	95.02%	94.89%	94.77%
Slashdot	89.94%	90.12%	89.65%
Wikipedia	88.63%	88.48%	88.75%

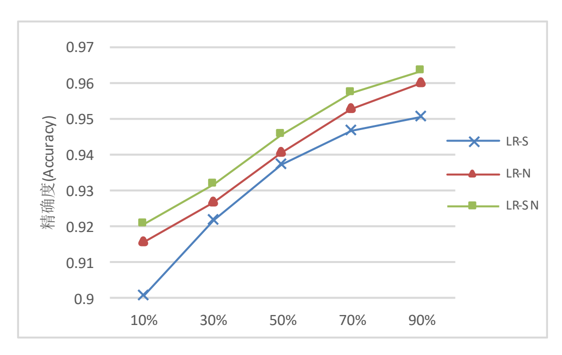
表 4　符号预测模型 lr-sn 的泛化能力（训练集占比 90%）

Table 4　Generalization ability of the sign prediction model LR-SN (training ratio 90%)

训练集	测试集		
	Epinion	Slashdot	Wikipedia
Epinion	96.31%	96.05%	95.74%
Slashdot	90.56%	90.81%	90.29%
Wikipedia	88.74%	88.42%	88.96%

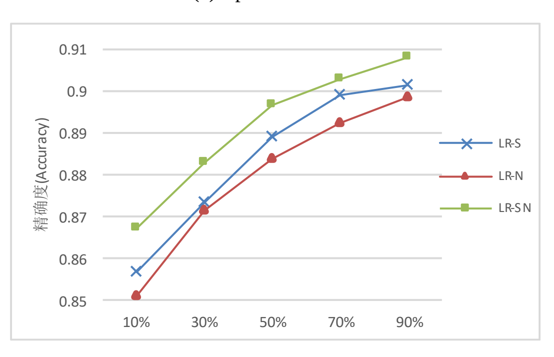
5.4　不同量化策略对边符号预测的影响

为了进一步验证节点地位以及相似性对边符号预测模型的影响，实验中还将节点相似性量化特征单独作为特征集输入，并运用逻辑回归进行符号预测，将该模型记为 LR-N。然后分别将 LR-S, LR-N 以及 LR-SN 三种模型进行比较。比对实验结果如图 4 所示。



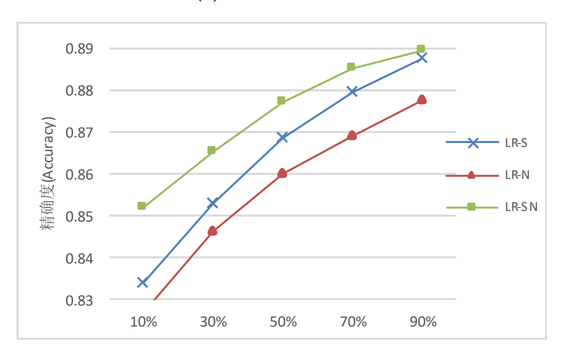
(a)Epinion 数据集

(a) Epinion dataset



(b)Slashdot 数据集

(b) Slashdot dataset



(c)Wikipedia 数据集

(c) Wikipedia dataset

图 4　不同量化策略对比结果

Fig. 4　Comparison of different quantitative strategies

通过分析图 4 可以得出以下结论：

a）从三个模型的对比结果可以看出，LR-SN 模型在三个数据集的预测效果均优于 LR-S 模型与 LR-N 模型。这一方面验证了将网络中的局部结构信息和全局结构信息融合有助于提高符号预测准确率；另一方面也说明了节点地位量化策略与节点相似性量化策略对符号预测的有效性。

b）LR-S 模型与 LR-N 模型在三个数据集中均获得了较高的预测效果，其中 LR-S 模型在 Slashdot 和 Wikipedia 数据集中表现更好，而 LR-N 模型在 Epinion 数据集中表现更好。分析其原因，可能是因为三大数据集中，Epinion 的聚类系数最高也最为密集，对于考虑局部信息的节点相似性其能获得的有效信息更多。而 Slashdot 和 Wikipedia 数据集较为稀疏，因此考虑全局信息的节点地位更占优势。

6　结束语

本文通过探索节点地位以及相似性二者与边符号属性之间的强相关性，运用逻辑回归方法 LR 实现了社交网络中的边符号预测问题。并针对二者分别建立基于节点地位的边符号预测模型 LR-S 以及结合节点地位与相似性的边符号预测模型 LR-SN，其中节点地位从全局角度量化符号属性相关特征，节点相似性从局部角度体现符号属性。三个真实网络数据集的实验结果表明，所提模型相比于现有基准方法，符号预测精确度显著提高，且具有一定的推广性。如何进一步探索符号属性相关的影响因素，并且使用更多的真实网络数据集来验证模型性能，是接下来的研究重点。

参考文献：

[1] 苏晓萍，宋玉蓉.符号网络的局部标注特征与预测方法 [J]. 智能系统学报，2018，13(3): 437-444. (Su Xiaoping, Song Yurong. Local signing features in signed networks and method of sign prediction [J]. Journal of Intelligent Systems, 2018, 13 (3): 437-444.)

[2] Tang Jiliang, Aggarwal C, Liu Huan. Recommendations in signed social networks [C]// Proc of the 25th International Conference on World Wide Web.2016: 31-40.

[3] Li Dong, Xu Zhiming, Chakraborty N, *et al*. Polarity related influence maximization in signed social networks [J]. PLoS One, 2014, 9(7): e102199.

[4] 王鹏，宋艳红，李松江，等. 针对行为特征的社交网络异常用户检测方法 [J]. 计算机应用，2017，37(S2): 219-224. (Wang Peng, Song Yanhong, Li Songjiang, *et al*. Detection method for abnormal user of social network based on behavior characteristics [J]. Journal of Computer Applications, 2017, 37 (S2): 219-224.)

[5] Heider F. Attitudes and cognitive organization [J]. Journal of Psychology, 1946, 21(1): 107-112.

[6] Cartwright D, Harary F. Structural balance: a generalization of Heider's theory [J]. Psychological Review, 1956, 63(5): 277-293.

[7] Leskovec J, Huttenlocher D, Kleinberg J. Predicting positive and negative links in online social networks [C]// Proc of the 19th International Conference on World Wide Web. New York: ACM Press, 2010: 641-650.

[8] Javari A, Mahdi J. Cluster-based collaborative filtering for sign prediction in social networks with positive and negative Links [J]. ACM Trans on Intelligent System and Technology, 2014, 5(2): 1-24.

[9] Symeonidis P, Tiakas E. Transitive node similarity: predicting and recommending links in signed social networks [J]. World Wide Web, 2014, 17(4): 743-776.

[10] Chiang K Y, Natarajan N, Tewari A, *et al*. Exploiting longer cycles for link prediction in signed network [C]//Proc of the 20th ACM International Conference on Information and Knowledge Management.

New York: ACM Press, 2011: 1157-1162.

[11] Matsuo Y, Yamamoto H. Community gravity: measuring bidirectional effects by trust and rating on online social networks [C]// Proc of the 18th Int Conf on World Wide Web. New York: ACM Press, 2009: 751-760.

[12] Zolfaghar K, Aghaie A. Mining trust and distrust relationships in social Web applications [C]// Proc of the 6th International Conference on Intelligent Computer Communication and Processing. Piscataway, NJ: IEEE Press, 2010: 73-80.

[13] Shahriari M,Jalili M.Ranking nodes in signed social networks [J]. Social Network Analysis and Mining, 2014, 4(1): 1-12.

[14] Leskovec J, Huttenlocher D, Kleinberg J. Signed networks in social media[C]//Proc of SIGCHI Conference on Human Factors in Computing Systems, New York: ACM Press, 2010: 1361-1370.

[15] 程苏琦，沈华伟，张国清，等. 符号网络研究综述 [J]. 软件学报, 2014, 25(1): 1-15. (Cheng Suqi, Shen Huawei, Zhang Guoqing,et al. Survey of signed networks research [J]. Journal of Software, 2014, 25 (1): 1-15.)